Single-Cell Sequencing of Mouse Thymocytes Reveals Mutational Landscape Shaped by Replication Errors, Mismatch Repair, and H3K36me3

Aska, Elli-Mari

2020-09-25

Aska , E-M , Dermadi , D & Kauppi , L 2020 , ' Single-Cell Sequencing of Mouse Thymocytes Reveals Mutational Landscape Shaped by Replication Errors, Mismatch Repair, and H3K36me3 ' , iScience , vol. 23 , no. 9 , 101452 . https://doi.org/10.1016/j.isci.2020.101452

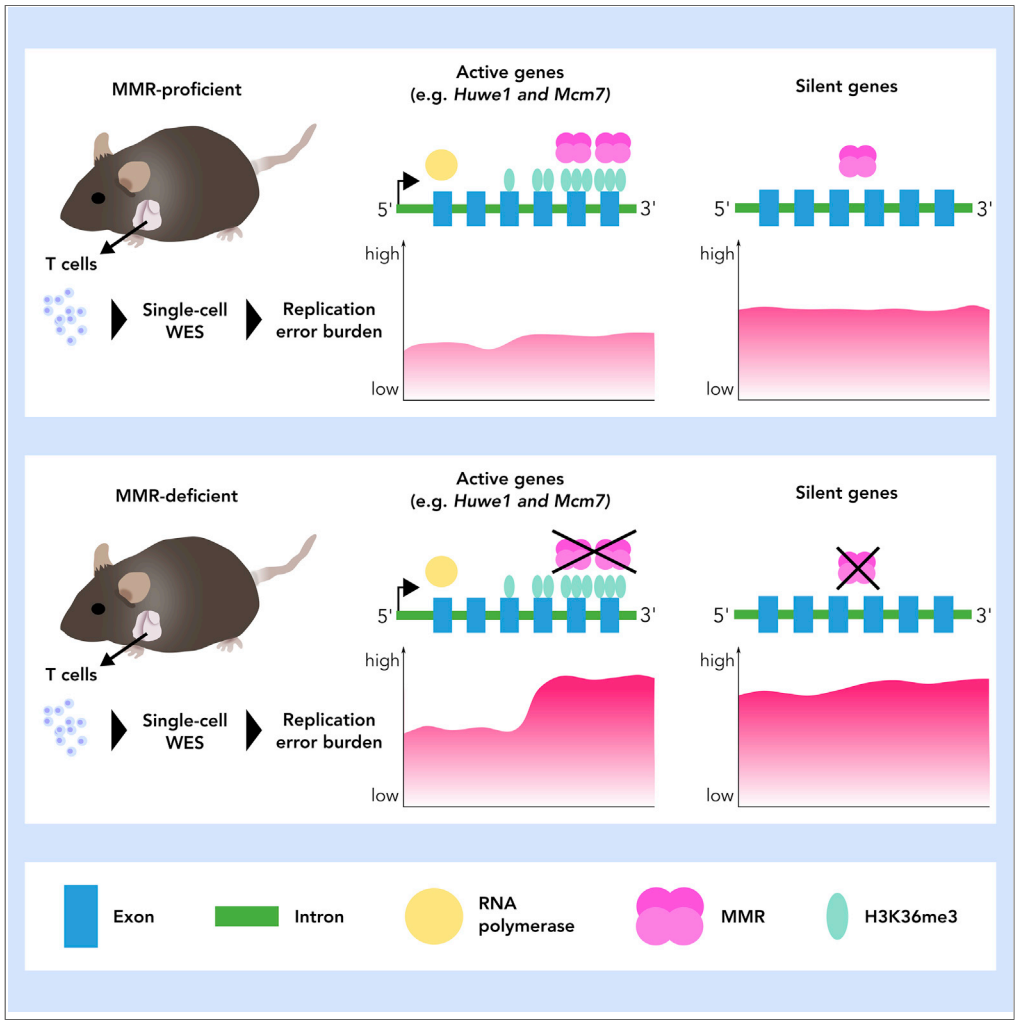
http://hdl.handle.net/10138/321119
https://doi.org/10.1016/j.isci.2020.101452

iScience

CellPress
OPEN ACCESS

Article

Single-Cell Sequencing of Mouse Thymocytes Reveals Mutational Landscape Shaped by Replication Errors, Mismatch Repair, and H3K36me3



Elli-Mari Aska, Denis Dermadi, Liisa Kauppi

elli.aska@helsinki.fi (E.A.)
ddermadi@stanford.edu (D.D.)
liisa.kauppi@helsinki.fi (L.K.)

HIGHLIGHTS

Mutational hotspots can be identified using single-cell sequencing in *Mlh1*$^{-/-}$ mice

Mcm7 and *Huwe1* genes represent mutational hotspots in non-malignant T cells

In vivo, 3′ exons of active genes enjoy MMR-mediated protection against mutations

Aska et al., iScience *23*, 101452
September 25, 2020
https://doi.org/10.1016/j.isci.2020.101452

iScience

CellPress
OPEN ACCESS

Article

Single-Cell Sequencing of Mouse Thymocytes Reveals Mutational Landscape Shaped by Replication Errors, Mismatch Repair, and H3K36me3

Elli-Mari Aska,[1,2,*] Denis Dermadi,[2,3,4,*] and Liisa Kauppi[1,2,5,*]

SUMMARY

DNA mismatch repair (MMR) corrects replication errors and is recruited by the histone mark H3K36me3, enriched in exons of transcriptionally active genes. To dissect *in vivo* the mutational landscape shaped by these processes, we employed single-cell exome sequencing on T cells of wild-type and MMR-deficient (*Mlh1*$^{-/-}$) mice. Within active genes, we uncovered a spatial bias in MMR efficiency: 3′ exons, often H3K36me3-enriched, acquire significantly fewer MMR-dependent mutations compared with 5′ exons. *Huwe1* and *Mcm7* genes, both active during lymphocyte development, stood out as mutational hotspots in MMR-deficient cells, demonstrating their intrinsic vulnerability to replication error in this cell type. Both genes are H3K36me3-enriched, which can explain MMR-mediated elimination of replication errors in wild-type cells. Thus, H3K36me3 can boost MMR in transcriptionally active regions, both locally and globally. This offers an attractive concept of thrifty MMR targeting, where critical genes in each cell type enjoy preferential shielding against *de novo* mutations.

INTRODUCTION

Maintaining genomic integrity during DNA replication is crucial for cellular homeostasis, especially in protein-coding regions. Occasionally, DNA replication errors occur, of which most, but not all, are corrected by the intrinsic proofreading activity of DNA polymerases (St Charles et al., 2015). DNA mismatch repair (MMR) corrects base-base mismatches and small insertion-deletion (indel) loops that have escaped proofreading and thereby protects the genome from replication-induced permanent mutations (Li, 2008). MMR initiates when the MSH2/MSH6 (MutSα) or MSH2/MSH3 (MutSβ) complex recognizes and binds DNA lesions, a step followed by recruitment of the MLH1/PMS2 (MutLα) complex that triggers the excision and repair of the mismatch (Lahue et al., 1989; Zhang et al., 2005).

MSH6 of MutSα can bind to trimethylated histone H3 lysine 36 (H3K36me3) and recruit the MMR machinery to chromatin (Li et al., 2013). H3K36me3 is found in exonic regions and enriched at the 3′ ends of transcribed genes (Kolasinska-Zwierz et al., 2009) and also in constitutive and facultative heterochromatin (Chantalat et al., 2011). Recently, H3K36me3 has been shown to also guide m^{6}A deposition to mRNA (Huang et al., 2019), which is known to affect mRNA stability and translation (Huang et al., 2018a; Wang et al., 2014, 2015). Genome-wide mutational analyses of MMR-deficient cell lines and tumors have shown that presence of H3K36me3 reduces local mutation rate (Supek and Lehner, 2015, 2017). Moreover, in tumors and cell lines, MMR operates more efficiently in H3K36me3-enriched exons compared with introns (Frigola et al., 2017), and in actively transcribed genes compared with silent genes, and lowers the mutation frequency in the 3′ ends of the genes (Huang et al., 2018b). Mutation signature of the error-prone polymerase η, which is part of the somatic hypermutation specific MMR pathway, is targeted to 3′ ends of genes via H3K36me3 in solid tumors (Supek and Lehner, 2017).

MMR deficiency has been extensively modeled in *Mlh1*$^{-/-}$ mice, which display high microsatellite instability (MSI) and increased tumor mortality (Baker et al., 1996; Edelmann et al., 1996, 1999; Prolla et al., 1998). Female *Mlh1*$^{-/-}$ mice frequently develop lymphomas, mainly thymic, whereas males tend to develop gastrointestinal tumors (Gladbach et al., 2019). MSI occurs owing to the propensity of microsatellites (short tandem repeat sequences) to undergo strand slippage during DNA replication, which in MMR-deficient cells leads to deletion or insertion mutations within repeats. Recently, analysis of genome-wide

[1]Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, 00290 Helsinki, Finland

[2]Department of Biochemistry and Developmental Biology, Faculty of Medicine, University of Helsinki, 00290 Helsinki, Finland

[3]Laboratory of Immunology and Vascular Biology, Department of Pathology, School of Medicine, Stanford University, Stanford, CA 94305, USA

[4]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, CA 94305, USA

[5]Lead Contact

*Correspondence: elli.aska@helsinki.fi (E.A.), ddermadi@stanford.edu (D.D.), liisa.kauppi@helsinki.fi (L.K.)

https://doi.org/10.1016/j.isci.2020.101452

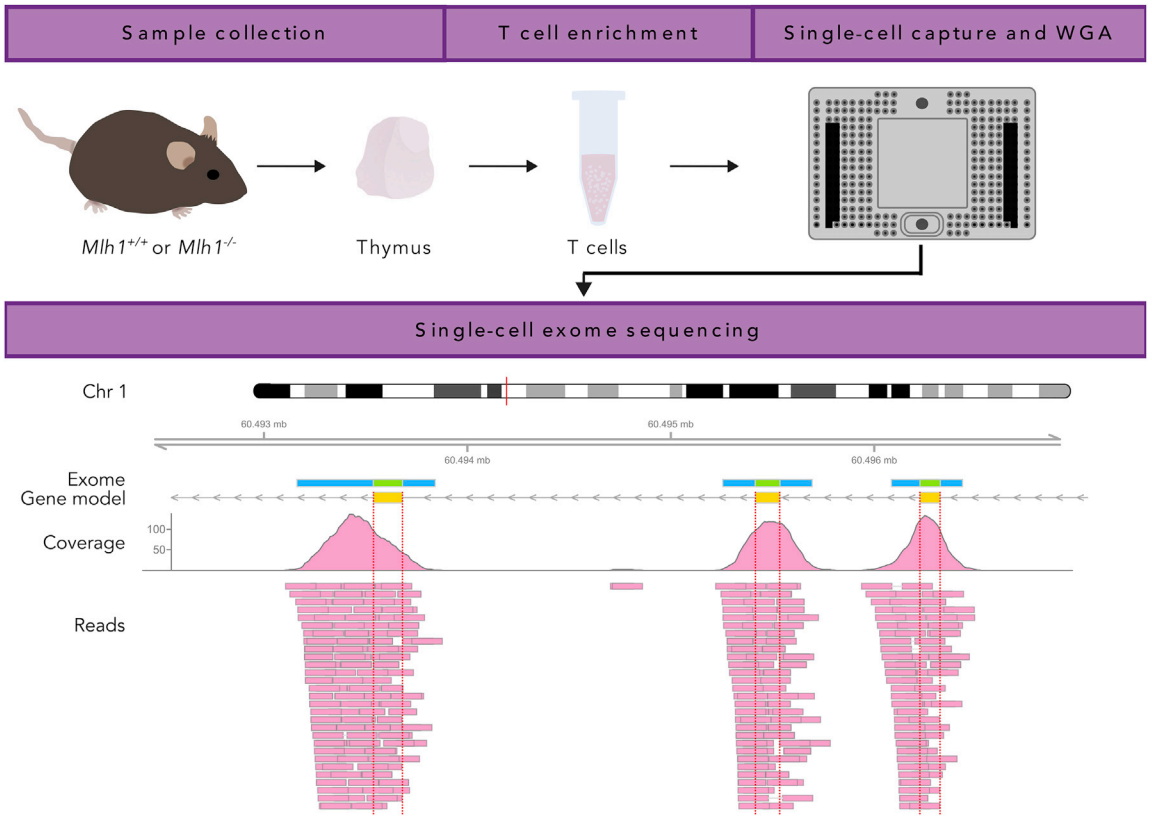


Figure 1. Whole-Exome Sequencing of Single T Cells: Experimental Overview

Thymi of *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ mice were dissected and used for enrichment of naive T cells, followed by single-cell capture, cell lysis, and whole-genome amplification in a Fluidigm C1. Amplified genomes were used for whole-exome sequencing (WES), and sequencing reads were analyzed for genetic variants. Shown is a read pileup and coverage of sample WT1-C26 in a ~5-kb-long region on chromosome 1 that contains three exons of *Raph1*. In addition to exons (green bar in exome panel), WES also partially covers non-coding regions adjacent to exons (blue bar in exome panel), enabling the comparison of mutation frequency between exonic and non-coding regions.

mutations in *Mlh1*$^{-/-}$ lymphomas revealed several putative drivers of tumorigenesis (Daino et al., 2019; Gladbach et al., 2019).

To delineate how the mutational landscape in normal mammalian cells is shaped *in vivo*, on one hand, by replication errors, and on the other hand, by H3K36me3-mediated MMR correction, we performed single-cell whole-exome sequencing (scWES) on T cells isolated from MMR-proficient (*Mlh1*$^{+/+}$) and MMR-deficient (*Mlh1*$^{-/-}$) mice. Comparison of mutation distribution and frequency between MMR-proficient and -deficient mice revealed *Huwe1* and *Mcm7* genes as mutational hotspots exclusive to *Mlh1*$^{-/-}$ cells, implying that these regions present an inherent challenge to faithful DNA replication in T cells. Both hotspots are located in H3K36me3-enriched regions and expressed during T cell development. Analysis of MMR-dependent mutations indicate that H3K36me3-enriched 3′ exons are more protected against transcription-associated replication errors.

RESULTS

Deletions Report on MMR-Dependent Mutations in Single-Cell Exome Sequencing

We isolated naive T cells from thymi of *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ mice, followed by single-cell capture and whole-genome amplification on the Fluidigm C1 system, and then, by whole-exome enrichment and sequencing (Figure 1). Previous studies have utilized single-cell DNA sequencing to study clonality and mutation profiles of human cancers and normal cells (Leung et al., 2017; Wu et al., 2017; Zhang et al., 2019; Pellegrino et al., 2018). To check whether T cells were drawn from a similar cell population in both genotypes, we analyzed the proportions of distinct developmental thymic T cell populations (double negative, double positive, TCR αβ single positive [CD4 or CD8], TCR γδ) (Shah and Zuniga-Pflucker, 2014) by

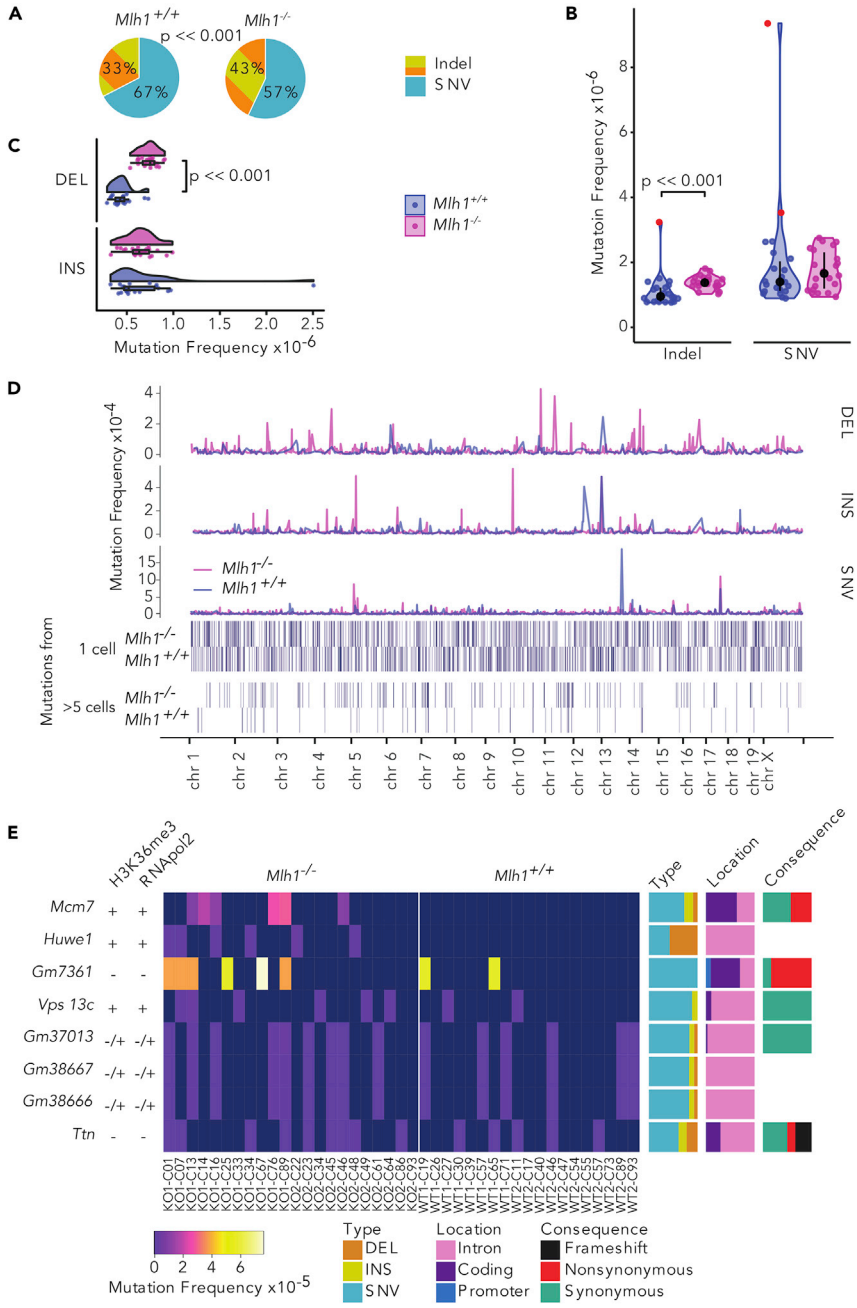


Figure 2. Global and Local Mutation Frequencies in Single T cells

(A–C) (A) *Mlh1*$^{-/-}$ T cells have an increased amount of indels out of total mutations in the whole exome compared with *Mlh1*$^{+/+}$ Tcells (p << 0.0001, Fisher's exact test). Global (B) indel and SNV frequencies and (C) deletion and insertion frequencies, in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ T cells.*Mlh1*$^{-/-}$ T cells have significantly higher indel, and especially deletion, frequencies than *Mlh1*$^{+/+}$ T cells (p << 0.001, two-tailed Mann-Whitney U test). Data in (B) is shown as boxplots together with kernel probability density and individual datapoints, and data in (C) is shown as median and interquartile range together with kernel probability density and individual datapoints. Outlier cells (see Transparent Methods) section ""Outlier cells in single-cell exomes) are marked with red color in (B).

(D) Mutation frequencies in 1-Mb windows across the mouse genome. *Mlh1*$^{-/-}$ T cells have multiple high local mutation peaks originating from only a single T cell.

(E) *Mcm7* and *Huwe1* are mutational hotspots in *Mlh1*$^{-/-}$ T cells. Columns are sorted by genotype and cell ID (outliers excluded), rows based on the average mutation frequency. *Mlh1*$^{+/+}$ cells have label WT and *Mlh1*$^{-/-}$ cells have label KO,

Figure 2. *Continued*
biological replicates are marked with 1 and 2. Each cell has a cell identifier that originates from the Fluidigm C1 plate capture site.
Bar plots on the right show proportions of mutation types, locations, and consequences in genes. Left-hand-side columns show positivity or negativity for RNApol2 and H3K36me3 peaks (See also Figure S3).

FACS. Cell frequencies of different thymic T cell populations between *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ mice were similar to each other (Figure S1), indicating no defect in normal T cell developmental progression in *Mlh1*$^{-/-}$ mice, and that T cells analyzed by scWES from *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ mice are drawn from similar thymic T cell populations. In both genotypes, the vast majority of cells were CD4+CD8+ double-positive T cells (67% for *Mlh1*$^{+/+}$ and 65% for *Mlh1*$^{-/-}$ mice, respectively, Figure S1).

We sequenced 56 single-cell exomes in total, from 28 *Mlh1*$^{-/-}$ and 28 *Mlh1*$^{+/+}$ T cells, to an average depth of 32X and coverage of 66% at depth ≥1X (Figures S2A and S2B). After excluding samples with low (<50%) coverage, 44 exomes (22 *Mlh1*$^{+/+}$ and 22 *Mlh1*$^{-/-}$ exomes) were further analyzed for genetic variants. All detected variants with annotations (Related to Transparent Methods sections "Variant calling and filtering" and "Mutation annotation") are listed in Table S2 titled "Annotated variants in single-cell exomes." Overall, *Mlh1*$^{-/-}$ T cells had increased percentage (odds ratio [OR] = 1.56, 95% confidence interval [CI] = 1.44–1.69, $p < 2.2 \times 10^{-16}$) and frequencies ($p = 5.487 \times 10^{-6}$, Figures 2A and 2B and Table S1) of indels when compared with *Mlh1*$^{+/+}$ T cells. Even though MMR deficiency increases also base substitutions (Meier et al., 2018), single nucleotide variant (SNV) frequencies between *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ did not differ significantly in our dataset (p = 0.127, Figure 2B and Table S1). Analyzing insertions and deletions separately revealed that *Mlh1*$^{-/-}$ T cells had significantly higher deletion ($p = 8.175 \times 10^{-12}$) but not insertion frequencies (p = 0.1801) than *Mlh1*$^{+/+}$ T cells (Figure 2C and Table S1). Taken together, deletions behaved in a genotype-dependent manner and thus represent MMR-dependent mutations.

Huwe1 and *Mcm7* Genes Are Mutational Hotspots in *Mlh1*$^{-/-}$ T Cells

Mlh1$^{-/-}$ cells provide a unique opportunity to reveal which chromosomal regions represent a particular challenge to the fidelity of the replication machinery, as any errors that are introduced will remain uncorrected by MMR. To identify such regions, we analyzed mutation frequencies in 1 Mb windows across single-cell exomes. On a megabase scale, local mutational frequencies were highly heterogeneous. The majority of the high-mutation-frequency peaks originated only from single T cells, and mutational hotspot windows shared between individual cells were sparse (Figure 2D). To establish whether any genes would emerge as MMR-dependent mutational hotspots, we scored all genes for mutations and asked which ones were mutated frequently in *Mlh1*$^{-/-}$ T cells (in more than 5 *Mlh1*$^{-/-}$ cells). Two genes, *Huwe1* and *Mcm7*, stood out with their high mutational frequencies, exclusive to *Mlh1*$^{-/-}$ single-cell exomes (Figure 2E). *Huwe1* encodes an E3 ubiquitin ligase, shown to regulate hematopoietic stem cell self-renewal and proliferation, and commitment to the lymphoid lineage (King et al., 2016). *Mcm7* encodes a component of the MCM2-7 complex that forms the core of the replicative helicase, responsible for unwinding DNA ahead of the replication fork (Deegan and Diffley, 2016). Both genes are positive for RNA polymerase 2 and H3K36me3 in the mouse thymus and expressed from hematopoietic stem cells all the way to thymic T cells (Figures 2E, S3A, and S3B).

We then compared the mutational hotspots in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ normal T cells (this study) and with those in *Mlh1*$^{-/-}$ mouse lymphomas (Kakinuma et al., 2007; Daino et al., 2019; Gladbach et al., 2019). Only one shared mutational hotspot gene was found: *Ttn*, a massive gene with 324 exons, was mutated in both *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ single-cell exomes (Figure 2E), in line with the findings of Daino et al. We did not identify any mutations in *Ikzf1*, previously reported as a mutational target gene in *Mlh1*-deficient T cell lymphomas (Daino et al., 2019; Kakinuma et al., 2007).

Other identified hotspot genes (*Gm7361*, *Vps13c*, *Gm37013*, *Gm38667*, *Gm38666*) were mutated in both *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ T cells and thus were not specific for *Mlh1* deficiency. All except *Vps13c* were negative or inconclusive for the presence of H3K36me3 and RNA polymerase 2, suggesting that these genes are not transcribed in mouse thymus (Figures 2E and S3A). *Gm37013*, *Gm38667*, and *Gm38666* are predicted genes and they physically overlap with each other on chromosome 18 (Figure S3A), which explains their identical mutational pattern.

Insertions and Deletions Accumulate Differently within Repeats in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ T Cells

Next, we analyzed the size distribution of detected indels in single-cell exomes. *Mlh1*$^{+/+}$ cells had more 1-nucleotide (nt) insertions than deletions, whereas this difference in *Mlh1*$^{-/-}$ T cells was evened out by increased 1-nt deletions (OR = 1.794, 95% CI = 1.531–2.101, p = 1.134 × 10^{-13}, Figure 3A). The same trend for 1-nt insertions as the dominant indel type in *Mlh1*$^{+/+}$ cells was observed in bulk T cell DNA samples from the same mice (Figure S4).

We then analyzed the sequence context of the detected indels. As expected, most deletions in *Mlh1*$^{-/-}$ cells occurred at mononucleotide microsatellites, whereas in *Mlh1*$^{+/+}$ cells, most deletions were found in non-microsatellite sequences (Figure 3B). When deletion counts were corrected for the number of base pairs of either microsatellite or non-microsatellite sequences, deletion frequencies were higher in microsatellites than in non-microsatellite sequences, regardless of MMR status (Figure 3C). This underscores the well-documented intrinsic propensity of microsatellites to slippage during replication. As expected, *Mlh1*$^{-/-}$ cells had significantly higher deletion frequencies in microsatellite sequences compared with *Mlh1*$^{+/+}$ cells (p = 9.505 × 10^{-13}, Figure 3C and Table S1). Insertion frequencies within repeats were more similar between *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ T cells, occurring especially in mononucleotide repeats (Figure 3D). *Mlh1*$^{-/-}$ cells had somewhat higher insertion frequencies in the context of microsatellite sequences when compared with *Mlh1*$^{+/+}$ cells (p = 0.039, Figure 3E and Table S1).

Exons Show a Decreased Burden of MMR-Dependent Mutations

Exome sequencing, despite its name, captures not only exons but also exon-adjacent, non-coding regions (3′ and 5′ UTR, promoter, or introns) (Figure 1) (Guo et al., 2012). This enabled us to ask whether *de novo* mutations accumulate differently in these two functionally distinct genic regions (exonic versus non-coding) in *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ cells.

No significant difference in SNV frequencies or insertions was observed in either exonic or non-coding regions in *Mlh1*$^{-/-}$ cells compared with *Mlh1*$^{+/+}$ cells (Figures 4A and 4B). In contrast, deletion frequencies increased in *Mlh1*$^{-/-}$ cells in non-coding regions compared with *Mlh1*$^{+/+}$ cells (p = 9.94 × 10^{-5}, Figure 4C and Table S1). Exonic deletion frequencies in *Mlh1*$^{-/-}$ cells did not differ from those observed in *Mlh1*$^{+/+}$ cells (Figure 4C), indicating that, in the absence of functional MMR, the integrity of coding regions is still maintained, likely by purifying selection, as suggested for MMR-deficient tumors by Kim et al., 2013. In conclusion, deletions, which we determined to be MMR-dependent mutations, increased more in non-coding regions adjacent to exons, as compared with exons themselves.

H3K36me3-Enriched Regions Are Depleted of MMR-Dependent Mutations

Results from large tumor datasets strongly indicate that exons have a decreased mutation burden due to H3K36me3-mediated MMR (Frigola et al., 2017), but evidence of this in normal cells and tissues *in vivo* is still lacking. To assess whether replication errors in transcribed genes are buffered by MMR by virtue of their H3K36me3 enrichment, we first analyzed H3K36me3 abundance in RNA polymerase 2 (RNApol2)-positive (RNApol2^{+}) and -negative (RNApol2^{-}) genes in thymus using publicly available ChIP-seq data (ENCODE Project Consortium, 2012; Sloan et al., 2016). Presence of RNA polymerase 2 in the promoter region is a strong indicator of transcriptional activity (Barski et al., 2007), and we used it to score genes as either active (RNApol2^{+}) or silent (RNApol2^{-}). H3K36me3 levels in RNApol2^{+} genes were higher than in RNApol2^{-} and peaked at the centers of the exons in these genes (Figure 5A), confirming that H3K36me3 is associated with transcriptional activity also in mouse thymus. However, not all RNApol2^{+} genes were positive for H3K36me3. Approximately 65% of RNApol2^{+} genes were also positive for H3K36me3, whereas 80% of H3K36me3-positive (H3K36me3^{+}) genes were positive for RNApol2 (Figure 5B).

We analyzed how small deletions (that is, MMR-dependent mutations) were distributed to exons and non-coding regions based on either RNApol2 or H3K36me3 status of genes. The proportion of exonic deletions over non-coding deletions was decreased in H3K36me3^{+} genes compared with H3K36me3-negative (H3K36me3^{-}) genes in *Mlh1*$^{+/+}$ (p = 0.018, OR = 0.44, 95% CI = 0.198–0.906) but not in *Mlh1*$^{-/-}$ T cells (p = 1, OR = 0.972, 95% CI = 0.542–1.694, Figures 5C and 5D). Lower exonic deletion burden in RNApol2^{+} genes was also observed in *Mlh1*$^{+/+}$ cells, similar to H3K36me3^{+} genes (p = 0.062, OR = 0.528, 95% C1 = 0.250–1.060, Figure 5C). The similar trends are not surprising, given the overlap between

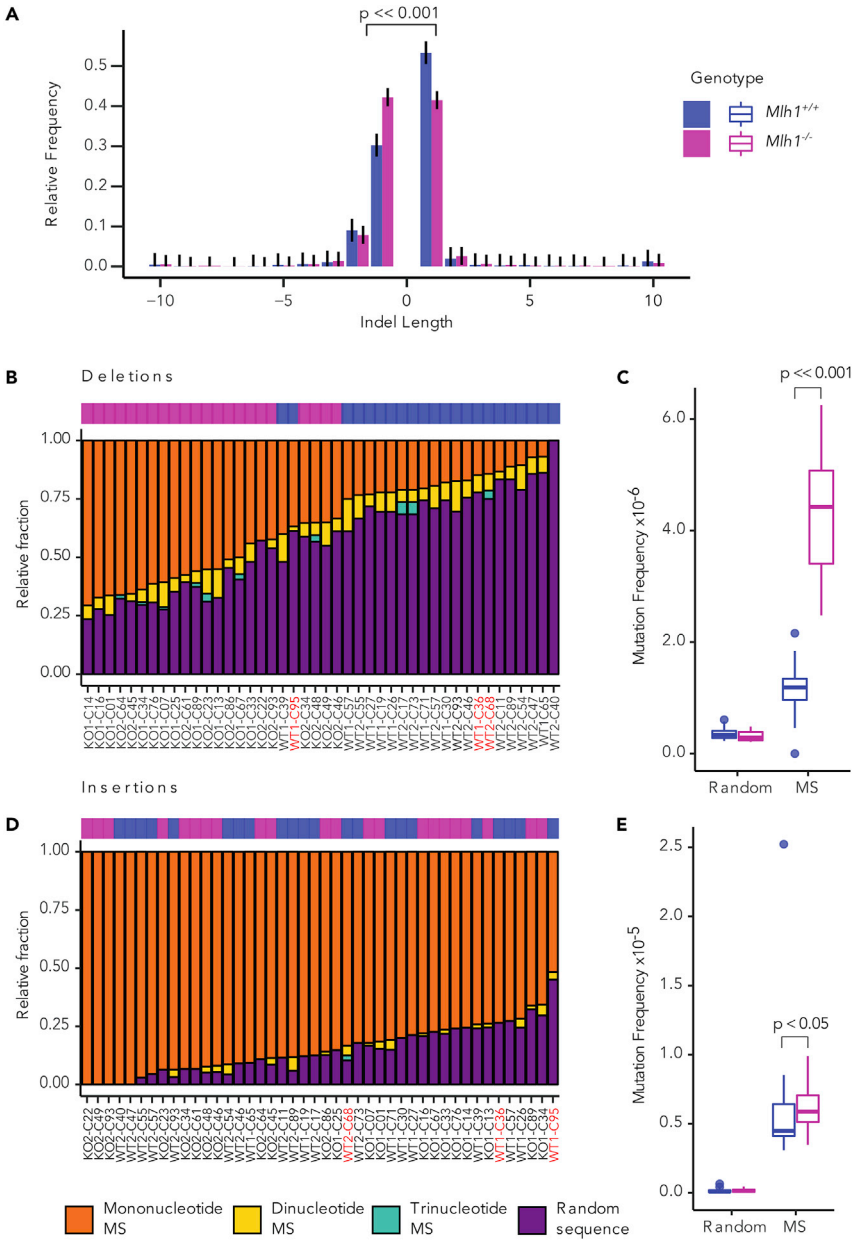


Figure 3. Small Deletions Report on MMR-Dependent Mutations in Mouse T cells

(A) Indel length distribution as relative frequencies with Sison and Glaz 95% multinomial confidence intervals in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ T cells. *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ cells have different ratios of 1-nt indels ($p << 0.001$, two-tailed Fisher's exact test). Indels of length ≥10 bp are binned together. See also Figure S4.

(B and C) (B) Relative and (C) normalized frequencies of deletions in microsatellites (MS) (mono-, di-, and trinucleotide repeats) and in non-microsatellite (random) sequence in single-cell samples.

(D and E) (D) Relative and (E) normalized frequencies of insertions in microsatellites (mono-, di-, and trinucleotide repeats) and in non-microsatellite (random) sequence in single-cell samples. Bar plots are ranked by descending mutation fraction within mononucleotide repeats. *Mlh1*$^{-/-}$ cells have a significantly higher deletion frequencies in microsatellites than *Mlh1*$^{+/+}$ ($p << 0.001$, two-tailed Mann-Whitney U test). Mutation frequencies are shown as boxplots. Outliers (see Transparent Methods section "Outlier cells in single-cell samples") are labeled with red in (B) and (D).

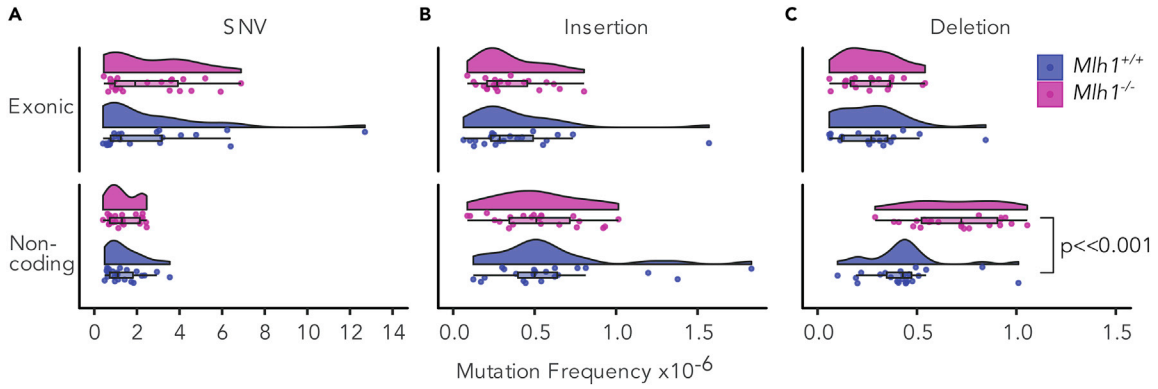


Figure 4. *Mlh1*$^{-/-}$ Cells Accumulate Mutations to Non-coding Regions of Genome

(A) SNV, (B) insertion, and (C) deletion frequencies in exonic and non-coding (3′ and 5′ UTRs, promoters, splice sites, introns) regions of the exome in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ T cells. *Mlh1*$^{-/-}$ T cells have significantly higher frequencies of non-coding deletions ($p<<0.001$, two-tailed Mann-Whitney U test). Data is shown as boxplots together with kernel probability density and individual datapoints

RNApol2^{+} and H3K36me3^{+} genes (Figure 5B). These results strongly support H3K36me3-guided, MMR-dependent protection of exons against genetic alterations.

The H3K36me3 mark is less abundant in 5′ exons, compared with 3′ exons of genes (Kolasinska-Zwierz et al., 2009; Frigola et al., 2017). To test whether local H3K36me3 levels affect the intra-genic distribution of mutations within genes *in vivo*, we compared deletion frequencies in the first and second exons (from here on referred to as 5′ exons) with those in the third to last exons (from here on referred to as 3′ exons), both in RNApol2^{+} and RNApol2^{-} genes. In RNApol2^{+} genes, H3K36me3 signal increased in 3′ exons compared with 5′ exons ($d = 0.335$, Figures 5E and S6A), whereas in RNApol2^{-} genes, there was no difference in H3K36me3 levels between 3′ and 5′ exons ($d = 0.002$, Figures 5F, S6A, and Table S1). In RNApol2^{+} genes, *Mlh1*$^{-/-}$ cells had higher deletion frequencies in 3′ exons (high in H3K36me3) compared with *Mlh1*$^{+/+}$ cells ($p = 4.57 \times 10^{-5}$, Figures 5E, S6B, and Table S1). In 5′ exons (low in H3K36me3), the difference in deletion frequencies between *Mlh1*$^{-/-}$ and *Mlh1*$^{+/+}$ was smaller, yet significant ($p = 0.016$, Figures 5E, S6B, and Table S1). *Mlh1*$^{+/+}$ cells also had somewhat increased deletion frequencies in the 3′ exons compared with 5′ exons ($p = 0.020$, Figures 5E, S6B, and Table S1). Sequencing coverage was similar between samples with or without mutations in the analyzed exons, except in the 5′ exons in RNApol2^{+} regions in *Mlh1*$^{+/+}$ cells ($p = 0.04$, Figure S5). Taken together, these results suggest that 3′ exons in transcriptionally active genes are more prone to acquiring replication-induced mutations compared with 5′ exons and that this effect is tempered by H3K36me3-guided MMR. No difference was observed in the deletion frequencies between *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells in RNApol2^{-} genes in 5′ exons ($p = 0.539$) or 3′ exons ($p = 0.296$, Figures 5F, S6B, and Table S1). *Mlh1*$^{-/-}$ cells, however, showed slightly higher deletion frequencies in 3′ exons compared with 5′ exons ($p = 0.049$, Figures 5F, S6B, and Table S1). H3K36me3^{-} exons in RNApol2^{-} genes accumulated mutations in similar frequencies in both *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells. We interpret this to mean that the MMR machinery does not operate efficiently in these regions even in wild-type cells. RNApol2^{+}, but not RNApol2^{-}, genes showed genotype-dependent spatial variability in deletion frequencies; thus transcriptional activity appears to affect accumulation and/or repair of replication errors.

DISCUSSION

Using single-cell sequencing of mouse thymic T cells, we uncovered how the exome-wide mutational landscape is shaped *in vivo* by replication errors, along with MMR-mediated error correction. We identify the *Huwe1* and *Mcm7* genes as novel mutational hotspots in normal *Mlh1*$^{-/-}$ thymic T cells. We further provide evidence for transcription-associated vulnerability to replication errors and for H3K36me3-guided MMR at 3′ exons of genes.

We show that scWES is a sensitive approach for unraveling signatures of replication errors and MMR activity. This is highlighted by the fact that we detected a substantial increase of deletions in *Mlh1*$^{-/-}$ T cells and found evidence of insertional bias in *Mlh1*$^{+/+}$ T cells. DNA polymerases tend to create more

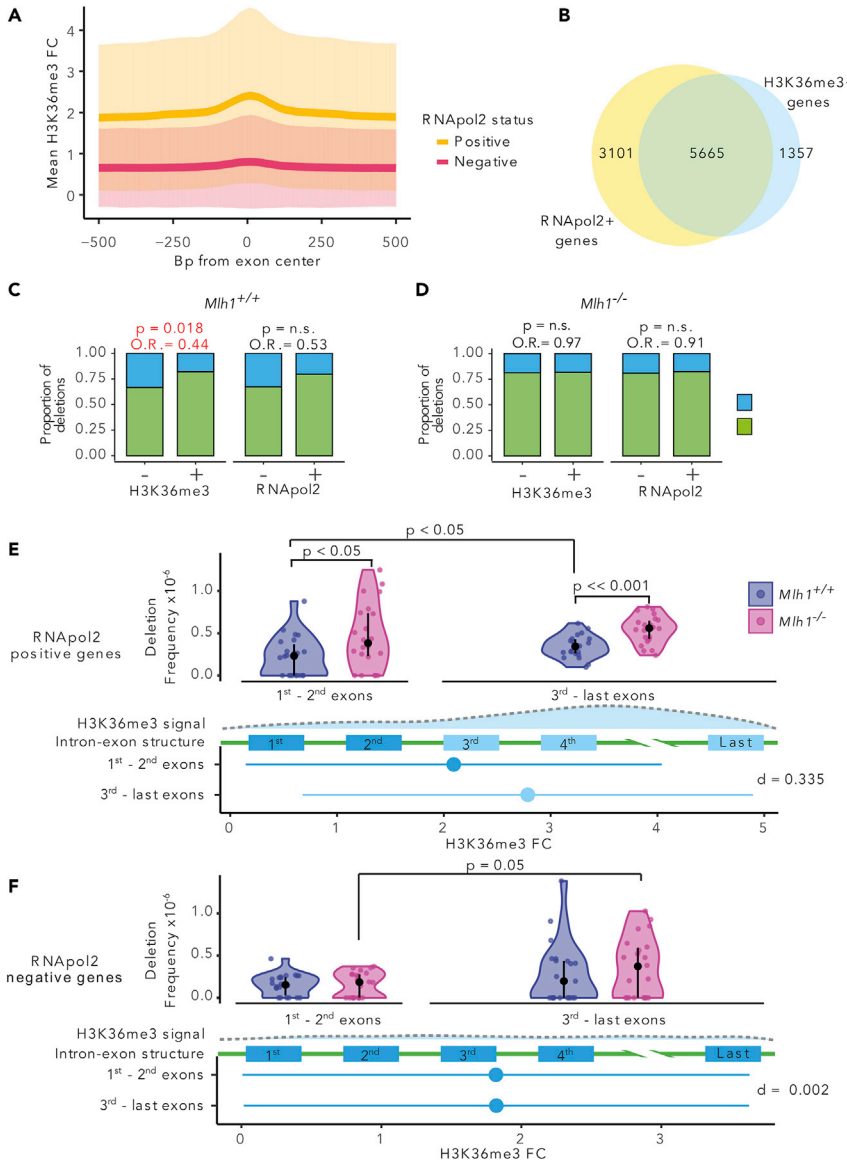


Figure 5. H3K36me3 Reduces the Amount of MMR-Dependent Mutations in Exons

(A) H3K36me3 fold change (FC) (mean ± SD) in 1,000-bp window around exon centers in RNApol2-positive and -negative genes.

(B) Venn diagram of RNApol2-positive (+) and H3K36me3-positive (+) gene counts. Proportions of small deletions in genes positive or negative for H3K36me3 and RNApol2 in (C) *Mlh1*$^{+/+}$ and (D) *Mlh1*$^{-/-}$ cells. Coding regions in genes positive for H3K36me3 have fewer deletions relative to H3K36me3-negative genes in *Mlh1*$^{+/+}$ cells ($p = 0.018$, OR = 0.44, two-tailed Fisher's exact test), but not in *Mlh1*$^{-/-}$ cells. Deletion frequencies in the first to second exons (5′ exons) and the third to last exons (3′ exons) in RNApol2 (E) -positive and (F) -negative genes. In RNApol2-positive genes, *Mlh1*$^{-/-}$ cells have higher deletion frequency especially in the third to last exons (high H3K36me3) than *Mlh1*$^{+/+}$ cells, and to lesser degree, in the first to second exons (low H3K36me3). The first panel shows the deletion frequencies in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells. data is shown as median and interquartile range together with kernel probability density and individual datapoints See also Figures S5 and S6. The second panel shows a schematic of H3K36me3 enrichment along a gene. The third panel shows a schematic of a gene structure. The fourth panel shows H3K36me3 signal as mean ± SD of FC in the first to second exons and third to last exons together with effect size as Cohen's d with Bessel's correction. Deletion frequencies were tested using two-tailed Mann-Whitney U test.

deletions than insertions, especially in repeat sequences (Baptiste et al., 2015; Kunkel, 1986; Kim et al., 2013; Lujan et al., 2015; Woerner et al., 2015; Garcia-Diaz and Kunkel, 2006), and in the absence of MMR (which is the situation in *Mlh1*$^{-/-}$ cells), one would expect to directly detect replication errors. Indeed, we observed a significant increase of small deletions in *Mlh1*$^{-/-}$ cells compared with *Mlh1*$^{+/+}$ cells. Taken together, we conclude that deletions reliably report on replication errors that would otherwise be repaired by MMR. In addition, we found that *Mlh1*$^{+/+}$ cells had more insertions than deletions. Increase in 1-nt insertions rather than deletions in *Mlh1*$^{+/+}$ cells has also been observed at unstable microsatellite loci in other MMR-proficient normal mouse tissues (Shrestha et al., 2019). Our findings are in line with the previously reported bias for MMR to correct deletion loops more efficiently than insertion loops, thereby creating an insertional bias at microsatellite sequences (Baptiste et al., 2013).

MMR-deficient cells (*Mlh1*$^{-/-}$) accumulate replication-induced errors with every cell division. Developing lymphocytes are particularly susceptible to replication errors because they undergo multiple rounds of proliferative expansions during development and maturation. Comparison of mutational frequencies in *Mlh1*$^{-/-}$ versus *Mlh1*$^{+/+}$ T cell exomes revealed two hotspots for replication errors, *Huwe1* and *Mcm7* genes. Mutations in *Mcm7* affected both exons and introns, whereas mutations in *Huwe1* were found exclusively in introns (Figure 2E). Exonic *Mcm7* mutations comprised both synonymous and non-synonymous mutations. Synonymous exonic *Mcm7* mutations, although they do not alter amino acid sequence, may still affect Mcm7 splicing regulatory sites or miRNA binding sites or cause changes in mRNA stability or translation efficiency. Intronic mutations may cause splicing defects, resulting in exon skipping or intron retention (Diederichs et al., 2016). A small fraction (1%–2%) of somatic mutations that alter amino acid sequence create neoepitopes that, when presented on the cell membrane, can provoke immune cell attack (Yamamoto et al., 2019). *Mlh1*-deficient mouse cancer cell lines have been shown to produce persistently neoantigens, both *in vitro* and *in vivo* (Germano et al., 2017). Neoantigenicity is unlikely, however, for MCM7 and HUWE1 that reside in the nucleus and/or cytosol, and thus, they lack the appropriate cellular localization to function as neoantigens. Because *Huwe1* and *Mcm7* are vulnerable to replication errors, we propose that over time, in *Mlh1*-deficient cells, damaging mutations will emerge in these genes, some with potentially tumorigenic effects. Indeed, deleterious mutations in *Huwe1* and *Mcm7* have been reported in *Mlh1*-deficient murine T cell lymphomas (Daino et al., 2019). The propensity of *Mcm7*, coding for an integral component of the replication machinery, to acquire deleterious mutations in MMR-deficient cells (Figure 2E) conceivably can further accelerate the accumulation of replication-associated errors, thereby adding insult to injury.

Both *Huwe1* and *Mcm7* are expressed in the T lymphocyte lineage and required for lymphocyte development. Shielding them from permanent mutations is likely important for cellular homeostasis and normal development, and *Huwe1* and *Mcm7* were in fact devoid of mutations in *Mlh1*$^{+/+}$ T cells. In the face of frequent replication errors, how is efficient targeting of MMR to these regions ensured in wild-type cells? Both *Huwe1* and *Mcm7* were enriched for H3K36me3 in the mouse thymus, and H3K36me3-mediated MMR has been shown to protect actively transcribed genes (Huang et al., 2018b). Thus, H3K36me3-mediated recruitment of MMR to these genes provides an explanation for efficient error correction in wild-type cells; in the absence of MMR, H3K36me3 no longer has a protective effect.

Also, at single-cell resolution, the protective effect of H3K36me3-mediated MMR on active genes appears to hold true more globally. In wild-type cells, coding regions in H3K36me3-enriched genes exhibited lower mutation frequencies, compared with coding regions in H3K36me3-depleted genes. This effect was abolished in MMR-deficient cells. Our results indicate that H3K36me3-mediated MMR preserves the integrity of active genes in normal tissues *in vivo*, similarly as shown previously for tumors and cell lines (Supek and Lehner, 2015; Frigola et al., 2017; Huang et al., 2018b).

Moreover, we provide *in vivo* evidence that 3′ ends of actively transcribed genes are more prone to replication-associated errors and that more efficient recruitment of MMR via H3K36me3 protects these regions, ensuring that most of these errors do not become permanent mutations. Head-on collisions of the replication and transcription machineries can cause indels and base substitutions and especially increase the deletion burden within 3′ ends (and to a lesser degree 5′ ends) of genes under active transcription (Sankar et al., 2016). In HeLa cells, mutation frequency has been shown to decrease toward the 3′ end of the gene body, as H3K36me3 increases, implying more efficient MMR-mediated repair in these regions (Huang et al., 2018b). SNVs also accumulate more to 3′ UTRs than to 5′ UTRs in aging B lymphocytes

(Zhang et al., 2019), in line with the notion that 3′ regions are in fact more prone to mutations. Efficient recruitment of the MMR machinery via H3K36me3 can shield against replication-induced errors specifically in transcribed genes, whose integrity is particularly important.

Here, we delineate the mutational landscape of T cells shaped by the status of DNA repair (functional versus impaired), dissected at the single-cell level in the context of H3K36me3. We provide evidence that, in normal thymocytes *in vivo*, MMR preferentially protects H3K36me3-positive genes and especially 3′ exons transcribed in T cell lineage, against accumulation of *de novo* mutations, providing an additional layer to the regional dynamics of H3K36me3-guided MMR. In addition, we identify *Huwe1* and *Mcm7* as novel mutational hotspots in (still phenotypically normal) *Mlh1*$^{-/-}$ T cells, both genes which are of importance during T cell development. Taken together, our results suggest an attractive concept of thrifty MMR targeting, where genes critical for the development of a given cell type and under mutational stress due to active transcription are preferentially shielded from acquiring deleterious mutations.

Limitations of the Study

- The number of sequenced single T cells in our study is limited. For a comprehensive view of mutational hotspots and mutation frequencies, more single cells should be sequenced.
- Owing to limited starting amount of DNA, single-cell genomes were amplified extensively in order to have enough material for sequencing. This amplification introduces *in vitro* artifacts, which affect the analysis of mutation frequencies and mutational features. Especially genuine *de novo* SNV mutations are expected to be masked by such artifacts.
- Our exomic dataset represents less than 2% of the whole mouse genome, and specifically the coding portion where *de novo* mutations are under highest natural selection. In order to understand how mutation frequency plays out in intergenic regions and the factors contributing to this dynamic, whole-genome sequencing should be conducted.

Resource Availability

Lead Contact

Any further queries and requests should be addressed to corresponding author and lead contact Liisa Kauppi (liisa.kauppi@helsinki.fi) or to corresponding authors Elli-Mari Aska (elli.aska@helsinki.fi) and Denis Dermadi (ddermadi@stanford.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Single-cell exome sequencing data generated and analyzed during the current study are deposited as raw reads in FASTQ format to SRA: PRJNA575619. The variants observed in single T cells supporting the conclusions of this article are provided with the article as Table S2 titled "Annotated variants in single-cell exomes" in xlsx file format. Publicly available H3K36me3 (ENCODE: ENCFF853BYO, ENCFF287DIJ) and RNApol2 (ENCODE: ENCFF119XEH) ChIPSeq data can be found from ENCODE (https://www.encodeproject.org) database.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101452.

ACKNOWLEDGMENTS

We are grateful to Fran Supek, Esa Pitkänen, Niko Välimäki, and Julia Casado for discussions and advice. We wish to acknowledge CSC – IT Center for Science, Finland for computing resources, the Functional Genomics Unit (University of Helsinki) for sequencing services, Minna Nyström (University of Helsinki) for providing mice, Jussi Taipale and Anna Vähärautio for access to Fluidigm C1 system, and Kul Shanker

Shrestha and Minna Tuominen for technical assistance. Assistance was also provided by the following core facilities: Laboratory Animal Center and Biomedicum Imaging Unit at University of Helsinki and Palo Alto Veterans Institute for Research (PAVIR) FACS Core.

E.A. is supported by a funded position in the Doctoral Program in Integrative Life Sciences, Doctoral School of Health, University of Helsinki, and ASLA-Fulbright Pre-Doctoral Fellowship 2018-2019. This work was supported by the Academy of Finland (grants 263870, 292789, 256996, 306026 to L.K.), the Sigrid Juséliuksen Säätiö (to L.K), and Emil Aaltonen Säätiö (to E.A.).

AUTHOR CONTRIBUTIONS

Conceptualization, D.D. and L.K.; Methodology, E.A. and D.D.; Formal Analysis, E.A. and D.D.; Resources, D.D. and L.K.; Data Curation, E.A.; Writing – Original Draft, E.A., D.D., and L.K.; Writing – Review & Editing, E.A., D.D., and L.K.; Visualization, E.A.; Supervision, D.D. and L.K.; Project Administration, D.D. and L.K.; Funding Acquisition, L.K.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: December 4, 2019
Revised: February 27, 2020
Accepted: August 10, 2020
Published: September 25, 2020

REFERENCES

Baker, S.M., Plug, A.W., Prolla, T.A., Bronner, C.E., Harris, A.C., Yao, X., Christie, D.M., Monell, C., Arnheim, N., Bradley, A., et al. (1996). Involvement of mouse Mlh1 in DNA mismatch repair and meiotic crossing over. Nat. Genet. *13*, 336–342.

Baptiste, B.A., Ananda, G., Strubczewski, N., Lutzkanin, A., Khoo, S.J., Srikanth, A., Kim, N., Makova, K.D., Krasilnikova, M.M., and Eckert, K.A. (2013). Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. G3 (Bethesda) *3*, 451–463.

Baptiste, B.A., Jacob, K.D., and Eckert, K.A. (2015). Genetic evidence that both dNTP-stabilized and strand slippage mechanisms may dictate DNA polymerase errors within mononucleotide microsatellites. DNA Repair (Amst) *29*, 91–100.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., WEI, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Chantalat, S., Depaux, A., Hery, P., Barral, S., Thuret, J.Y., Dimitrov, S., and Gerard, M. (2011). Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. Genome Res. *21*, 1426–1437.

Daino, K., Ishikawa, A., Suga, T., Amasaki, Y., Kodama, Y., Shang, Y., Hirano-Sakairi, S., Nishimura, M., Nakata, A., Yoshida, M., et al. (2019). Mutational landscape of T-cell lymphoma in mice lacking the DNA mismatch repair gene Mlh1: no synergism with ionizing radiation. Carcinogenesis *40*, 216–224.

Deegan, T.D., and Diffley, J.F. (2016). MCM: one ring to rule them all. Curr. Opin. Struct. Biol. *37*, 145–151.

Diederichs, S., Bartsch, L., Berkmann, J.C., Frose, K., Heitmann, J., Hoppe, C., Iggena, D., Jazmati, D., Karschnia, P., Linsenmeier, M., et al. (2016). The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Mol. Med. *8*, 442–457.

Edelmann, W., Cohen, P.E., Kane, M., Lau, K., Morrow, B., Bennett, S., Umar, A., Kunkel, T., Cattoretti, G., Chaganti, R., et al. (1996). Meiotic pachytene arrest in MLH1-deficient mice. Cell *85*, 1125–1134.

Edelmann, W., Yang, K., Kuraguchi, M., Heyer, J., Lia, M., Kneitz, B., Fan, K., Brown, A.M., Lipkin, M., and Kucherlapati, R. (1999). Tumorigenesis in Mlh1 and Mlh1Apc1638N mutant mice. Cancer Res. *59*, 1301–1307.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Frigola, J., Sabarinathan, R., Mularoni, L., Muinos, F., Gonzalez-Perez, A., and Lopez-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. Nat. Genet. *49*, 1684–1692.

Garcia-Diaz, M., and Kunkel, T.A. (2006). Mechanism of a genetic glissando: structural biology of indel mutations. Trends Biochem. Sci. *31*, 206–214.

Germano, G., Lamba, S., Rospo, G., Barault, L., Magri, A., Maione, F., Russo, M., Crisafulli, G., Bartolini, A., Lerda, G., et al. (2017). Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. Nature *552*, 116–120.

Gladbach, Y.S., Wiegele, L., Hamed, M., Merkenschlager, A.M., Fuellen, G., Junghanss, C., and Maletzki, C. (2019). Unraveling the heterogeneous mutational signature of spontaneously developing tumors in MLH1(-/-) mice. Cancers (Basel) *11*, 1485.

Guo, Y., Long, J., He, J., Li, C.I., Cai, Q., Shu, X.O., Zheng, W., and Li, C. (2012). Exome sequencing generates high quality data in non-target regions. BMC Genomics *13*, 194.

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B.S., Mesquita, A., Liu, C., Yuan, C.L., et al. (2018a). Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat. Cell Biol. *20*, 285–295.

Huang, Y., Gu, L., and Li, G.M. (2018b). H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. J. Biol. Chem. *293*, 7811–7823.

Huang, H., Weng, H., Zhou, K., Wu, T., Zhao, B.S., Sun, M., Chen, Z., Deng, X., Xiao, G., Auer, F., et al. (2019). Histone H3 trimethylation at lysine 36 guides m(6)A RNA modification co-transcriptionally. Nature *567*, 414–419.

Kakinuma, S., Kodama, Y., Amasaki, Y., Yi, S., Tokairin, Y., Arai, M., Nishimura, M., Monobe, M., Kojima, S., and Shimada, Y. (2007). Ikaros is a mutational target for lymphomagenesis in Mlh1-deficient mice. Oncogene *26*, 2945–2949.

Kim, T.M., Laird, P.W., and Park, P.J. (2013). The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell *155*, 858–868.

King, B., Boccalatte, F., Moran-Crusio, K., Wolf, E., Wang, J., Kayembe, C., Lazaris, C., Yu, X., Aranda-Orgilles, B., Lasorella, A., and Aifantis, I. (2016). The ubiquitin ligase Huwe1 regulates the maintenance and lymphoid commitment of hematopoietic stem cells. Nat. Immunol. *17*, 1312–1321.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. *41*, 376–381.

Kunkel, T.A. (1986). Frameshift mutagenesis by eucaryotic DNA polymerases in vitro. J. Biol. Chem. *261*, 13581–13587.

Lahue, R.S., Au, K.G., and Modrich, P. (1989). DNA mismatch correction in a defined system. Science *245*, 160–164.

Leung, M.L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S., and Navin, N.E. (2017). Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res. *27*, 1287–1299.

Li, G.M. (2008). Mechanisms and functions of DNA mismatch repair. Cell Res. *18*, 85–98.

Li, F., Mao, G., Tong, D., Huang, J., Gu, L., Yang, W., and Li, G.M. (2013). The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSalpha. Cell *153*, 590–600.

Lujan, S.A., Clark, A.B., and Kunkel, T.A. (2015). Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects. Nucleic Acids Res. *43*, 4067–4074.

Meier, B., Volkova, N.V., Hong, Y., Schofield, P., Campbell, P.J., Gerstung, M., and Gartner, A. (2018). Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. Genome Res. *28*, 666–675.

Pellegrino, M., Sciambi, A., Treusch, S., Durruthy-Durruthy, R., Gokhale, K., Jacob, J., Chen, T.X., Geis, J.A., Oldham, W., Matthews, J., et al. (2018). High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. Genome Res. *28*, 1345–1352.

Prolla, T.A., Baker, S.M., Harris, A.C., Tsao, J.L., Yao, X., Bronner, C.E., Zheng, B., Gordon, M., Reneker, J., Arnheim, N., et al. (1998). Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DNA mismatch repair. Nat. Genet. *18*, 276–279.

Sankar, T.S., Wastuwidyaningtyas, B.D., Dong, Y., Lewis, S.A., and Wang, J.D. (2016). The nature of mutations induced by replication-transcription collisions. Nature *535*, 178–181.

Shah, D.K., and Zuniga-Pflucker, J.C. (2014). An overview of the intrathymic intricacies of T cell development. J. Immunol. *192*, 4017–4023.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.

St Charles, J.A., Liberti, S.E., Williams, J.S., Lujan, S.A., and Kunkel, T.A. (2015). Quantifying the contributions of base selectivity, proofreading and mismatch repair to nuclear DNA replication in *Saccharomyces cerevisiae*. DNA Repair (Amst) *31*, 41–51.

Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature *521*, 81–84.

Supek, F., and Lehner, B. (2017). Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. Cell *170*, 534–547.e23.

Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

Wang, X., Zhao, B.S., Roundtree, I.A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N(6)-methyladenosine modulates messenger RNA translation efficiency. Cell *161*, 1388–1399.

Woerner, S.M., Tosti, E., Yuan, Y.P., Kloor, M., Bork, P., Edelmann, W., and Gebert, J. (2015). Detection of coding microsatellite frameshift mutations in DNA mismatch repair-deficient mouse intestinal tumors. Mol. Carcinog *54*, 1376–1386.

Wu, H., Zhang, X.Y., Hu, Z., Hou, Q., Zhang, H., Li, Y., Li, S., Yue, J., Jiang, Z., Weissman, S.M., et al. (2017). Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. Oncogene *36*, 2857–2867.

Yamamoto, T.N., Kishton, R.J., and Restifo, N.P. (2019). Developing neoantigen-targeted T cell-based treatments for solid tumors. Nat. Med. *25*, 1488–1499.

Zhang, Y., Yuan, F., Presnell, S.R., Tian, K., Gao, Y., Tomkinson, A.E., GU, L., and LI, G.M. (2005). Reconstitution of 5'-directed human mismatch repair in a purified system. Cell *122*, 693–705.

Zhang, L., Dong, X., Lee, M., Maslov, A.Y., Wang, T., and Vijg, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. Proc. Natl. Acad. Sci. U S A *116*, 9014–9019.

Shrestha, K., Aska, E., Tuominen, M. & Kauppi, L. (2019). Mlh1 haploinsufficiency induces microsatellite instability specifically in intestine, https://doi.org/10.1101/652198.

iScience, Volume *23*

Supplemental Information

Single-Cell Sequencing of Mouse Thymocytes Reveals Mutational Landscape Shaped by Replication Errors, Mismatch Repair, and H3K36me3

Elli-Mari Aska, Denis Dermadi, and Liisa Kauppi

SUPPLEMENTAL INFORMATION

Transparent Methods

Mice

Two female *Mlh1*$^{-/-}$ (Edelmann et al., 1996) and two of their *Mlh1*$^{+/+}$ female littermates, age 12 weeks, were used for the single-cell whole exome sequencing study. All animal experiments were performed following national and institutional guidelines (the National Animal Experiment Board in Finland and the Laboratory Animal Centre of the University of Helsinki) under animal license number ESAVI/1253/04.10.07/2016.

Enrichment of thymic T cells

Mice were euthanized by carbon dioxide inhalation, followed by cervical dislocation. Thymi were collected in ice-cold DMEM (Gibco cat: 11960-044) and visually inspected for any macroscopic anomalies. Whole thymi were homogenized for an enrichment of naïve T cells using a commercially available kit according to manufacturer's instructions (Invitrogen, cat:11413D).

Single-cell capture and whole genome amplification

Enriched T cells were prepared for single-cell capture and whole-genome amplification in Fluidigm C1 system according to manufacturer's protocol (Fluidigm cat: 100-7357). Single T cells were captured using an IFC 5-10 µm capture plate (Fluidigm cat: 100-5762) and imaged using Nikon Eclipse Ti-E microscope with Hamamatsu Flash 4.0 V2 scientific CMOS detector. After confirming the capture by microscopy, cell lysis and whole-genome amplification steps were carried out in Fluidigm C1 system using illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences cat: 25-6600-30). DNA concentrations of amplified single-cell genomes were determined using either a Qubit dsDNA HS Assay kit (Invitrogen cat:Q32854) with Qubit Fluorometer (1.27) or QuantiFluor dsDNA System (Promega cat:E2670) with Quantus Fluorometer (2.24). Fragment size and integrity of amplified single-cell genomes were analyzed using Bioanalyzer High Sensitivity DNA Assay (Agilent) with Agilent Bioanalyzer 2100 (2100 Expert B.02.08.S648 SR3) or TapeStation Genomic DNA ScreenTape (Agilent) with TapeStation 4200 (TapeStation Analysis Software A.02.021 SR1) at the Biomedicum Functional Genomics Unit, Helsinki. Samples with the highest density of fragments around ~10 kb were chosen for sequencing based on visual inspection of the fragment size distributions.

Library preparation and sequencing

Agilent SureSelectXT Mouse All Exon 49.6Mb capture was used for exome enrichment and to prepare multiplexed libraries for Illumina. Samples were sequenced using Illumina NextSeq 500 with mid output reagents as paired-end 150 bp reads. In total, we sequenced 56 single T cell exomes in three batches, each batch consisting of single-cell samples with a genotype-matched bulk DNA sample (= whole genome amplified cell suspension, n=3/genotype, biological replicates 1 and 2, and technical replicate for biological replicate 1). Sequencing was performed by the Biomedicum Functional Genomics Unit, Helsinki.

Sequence alignment

Sequence alignment and variant calling workflow was adapted from Leung et al. (Leung et al., 2016). Paired-end reads were aligned to the Dec. 2011 (GRCm38/mm10) assembly of the mouse genome using bowtie2 (2.3.4) (Langmead and Salzberg, 2012) with --local mode. Aligned reads were then sorted, merged, and marked for duplicates using SAMtools (1.4) (Li et al., 2009) and Picard (2.13.2). Reads were re-aligned around indels using GATK (3.8-0-ge9d806836) (McKenna et al., 2010), followed by removal of reads with low mapping quality (MQ < 40) using SAMtools. Sequencing metrics (average depth and coverage) were calculated using SAMtools, BEDtools (2.26.0) (Quinlan and Hall, 2010) and R (3.5.0). Samples that had coverage less than 50% at depth ≥1X were excluded from subsequent analyses (**Fig. S1B**).

Variant calling and filtering

Variants within the exome capture region + 100 bp interval padding were called using GATK HaplotypeCaller in -ERC GVCF mode, followed by joint calling with GenotypeGVCFs. Samples (single-cell and bulk DNA) from the same genotype (*Mlh1*$^{+/+}$ or *Mlh1*$^{-/-}$) were analyzed together. Variant score recalibration was done separately to indels and SNVs using GATK SelectVariants and VariantRecalibration and applied at 99.0 sensitivity level using ApplyRecalibration. Variant sets used to build the recalibration model for SNVs were dbSNP (build 150) (Sherry et al., 2001), Mouse Genomes Project SNP Release Version 5 (Keane et al., 2011), and bulk SNV set (see below), and for indels, dbSNP (build 150), Mouse Genomes Project indel Release Version 5, and bulk indel set (see Transparent Methods section "High confidence bulk indel and SNV training set construction"). After variant score recalibration, all variants that had genotype quality <20, depth <6 and heterozygous genotypes allelic depth <0.333 were filtered out. Clustered SNVs (>3 SNVs / 10 bp) were filtered out to eliminate false positive SNVs caused by poor alignment around indels. Variants found in both *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ samples (germline mutations), homozygous mutations (insufficient whole-genome amplification) and variants found in the 129P2 OlaHsd strain were excluded from all subsequent analyses (mice with disrupted *Mlh1* were originally created using 129/Ola derived embryonic stem cells that were injected to C57BL/6 mice (Edelmann et al., 1996)). Filtering was done using GATK VariantFiltration, Picard FilterVcf, and R package *VariantAnnotation* (1.26.1) (Obenchain et al., 2014). All variants detected with annotations (See also Transparent Methods section "Mutation annotation") are listed in Supplemental Table S2 titled "Annotated variants in single-cell exomes".

High confidence bulk indel and SNV training set construction

High confidence bulk DNA SNV and indel training sets for variant score recalibration were constructed from the raw variants discovered in bulk DNA samples (both *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$) by including the variants that passed the following filters: ReadPosRankSum > -1.9, QD > 5.0, SOR > 1.5 for indels and SNVs, and for SNVs only: MQRankSum > -1.9. Variants that did not have a genotype (= insufficient sequencing coverage) across all bulk samples (n=3/genotype) were removed from the reference bulk set.

Mutation annotation

Mutations were annotated (gene, genic location, mutation consequence) using R package *VariantAnnotation* function *locateVariants* with *AllVariants* option and *predictCoding*. The UCSC KnownGene track from *TxDb.Mmusculus.UCSC.mm10.knownGene* (3.4.0) was used as the gene model. We considered mutations that fall within CDS regions to be exonic, and those that fall within 5′ untranslated region (UTR), 3′ UTR, splice site, intron or promoter to be non-coding. For the analysis of exonic and non-coding indels (**Fig. 4A-C** and **Fig. 5C-D**), we included mutations in genes with only one transcript to avoid having multiple locations within one gene for one mutation. In the mutation hotspot analysis (**Fig. 2E**), all possible transcript variants were analyzed.

Regions with transcriptional activity and enriched with H3K36me3 in mouse exome

RNApol2 (ENCODE: ENCFF119XEH) and H3K36me3 (ENCODE: ENCFF853BYO) ChIP-seq peak coordinates for mouse thymus were downloaded as BED files from ENCODE (Consortium, 2012, Sloan et al., 2016). We used UCSC knownGene track to define the genomic coordinates of genes. Genes that overlapped or were within 100 bp of the ChIP-seq peak coordinates were defined positive for that feature. Genes positive for H3K36me3 or RNApol2 peaks were defined separately.

H3K36me3 signal in genes

H3K36me3 data (ENCODE: ENCFF287DIJ) for mouse thymus was downloaded as a BigWig file containing fold-change (FC) of ChIP reads over background reads from ENCODE. Mean H3K36me3 FC ± standard deviation (s.d.) in each position (meaning, each *base* gets a mean H3K36me3 FC value) 500 bases up- and downstream from the exome capture fragment centers was calculated for RNApol2- positive and -negative genes (see Transparent Methods section "Regions with transcriptional activity and enriched with H3K36me3 in mouse exome"). Mean H3K36me3 FC ± s.d. in 5′ and 3′ exons (meaning, each *region* gets a mean H3K36me3 FC value) were calculated for RNApol2-positive and -negative genes.

Microsatellites in mouse exome

Mono-, di-, and trinucleotide repeats in mouse exome were detected using STR-FM (Galaxy version 1.0.0) (Fungtammasan et al., 2015) in Galaxy at usegalaxy.org (Afgan et al., 2018). R package *BSgenome.Mmusculus.UCSC.mm10* (1.4.0) was used to convert BED file containing genomic coordinates of variant call regions into FASTA format. Mono-, di-, and trinucleotide repeats were detected from the FASTA file in separate runs using motif sizes 1, 2, and 3, no partial motifs allowed, and minimum repeat unit counts were 4 (minimum length 4 bp) in mononucleotide repeat detection and 3 in dinucleotide (minimum length 6 bp) and trinucleotide (minimum length 9 bp) repeat detections. Non-microsatellite associated regions were defined as those that were not defined as mono-, di- nor trinucleotide repeats.

Microsatellite associated indels in single-cells

Sequence 100 bp up- and downstream of detected indel start coordinates were extracted from the mouse reference genome mm10 (*BSgenome.Mmusculus.UCSC.mm10*) in FASTA format and analyzed for mono-, di- and trinucleotide repeats as described above (See Transparent Methods section "Microsatellites in mouse exome"). Indels were marked microsatellite-associated if the indel start coordinate and microsatellite start coordinate were the same. Indels found not to be within mono-, di- or trinucleotide repeat were labelled as non- microsatellite associated (random) indels.

Mutation frequencies in single T cells

Global indel and SNV frequencies in the variant call region were calculated for each single-cell and reported as mutations/base. Mutation frequency was calculated as: $frq = n/(cov*2)$, where n is the number of mutations, *cov* is the number of high-quality base pairs (MQ > 40, DP > 6). Similarly, frequencies in different genomic regions (exonic, non-coding, microsatellites, 3′ exons, 5′ exons) were calculated by first counting the number of mutations in each region and dividing it by the coverage of that particular region.

Mutation frequencies in 1 Mb windows

Local mutation frequencies in 1 Mb windows were calculated by first dividing the genome into 1 Mb windows, then calculating the coverage of the variant call region (exome capture + 100 bp padding) in each window. Next, the number of SNVs, deletions, and insertions per genotype ($Mlh1^{+/+}$ or $Mlh1^{-/-}$) was counted in each window. Mutation frequency for $Mlh1^{+/+}$ and $Mlh1^{-/-}$ groups was then calculated by dividing the number of observed mutations in each window by the coverage (*cov*2*) of variant call region in that window.

Mutation hotspot analysis

We analyzed all genes for mutations in $Mlh1^{+/+}$ and $Mlh1^{-/-}$ T cells. For each sample, we counted the number of mutations per gene. These numbers were then normalized by the coverage (*cov*2*) of the gene in each sample. A gene was considered to be a hotspot if it was mutated in more than 5 $Mlh1^{-/-}$ T cells.

Outlier cells in single-cell samples

Cells that had indel or SNV frequency higher or lower than 1.5 * interquartile range in matching genotype were labelled as outliers and removed from all the subsequent statistical test. Outliers are shown in the plots, unless mentioned otherwise, and indicated in **Figs. 2B**, **3B** and **3D**.

MMR dependent mutation frequencies in 5′ and 3′ exons

To analyze mutation frequencies and H3K36me3 signal in 5′ exons (1st to 2nd exons) and 3′ exons (3rd to last exons), we took UCSC knownGene transcripts, excluded genes that overlap each other, and collapsed transcripts gene-wise to create one exon-intron-structure for each gene. 100 bp padding was added to each exon. Only genes with 4 or more exons were considered and exons 1-2 were marked as 5′ exons and exons 3rd to last were marked as 3′ exons. Genes that were in or within 100 bp of RNApol2 peak coordinates were marked as RNApol2 positive. Number of deletions in 5′ and 3′ exons in each single-cell were counted and then divided by the coverage (*cov*2*) of either 3′ or 5′ exons in that single-cell sample.

General R packages

R version 3.5.0 was used to analyze the data. *VariantAnnotation* package was used for VCF file manipulation, *rtracklayer* (1.40.3) (Lawrence et al., 2009) package for reading BED and BigWig files, and *GenomicRanges* (1.32.6) (Lawrence et al., 2013) package for handling genomic coordinates in R environment. Figures and general data manipulation were done using *ggplot2* (3.00.0), *gplots* (3.00.1), *Gviz* (1.24.0), *grid* (3.5.0), *viridis* (0.5.1), *dplyr* (0.7.6), *plyr* (1.8.4), *reshape2* (1.4.3), *tidyr* (0.8.2), *VennDiagram* (1.6.20), and *Hmisc* (4.1-1).

Statistical analysis

All tests were calculated using 22 *Mlh1*$^{-/-}$ T cells and 19 *Mlh1*$^{+/+}$ T cells, except in the **Fig. 2A**, where all single cell samples were included (22 *Mlh1*$^{-/-}$ T cells and 22 *Mlh1*$^{+/+}$ T cells). All mutation frequencies are reported as median (mdn) and interquartile range (iqr) (**Table S1**) and tested using two-tailed Mann-Whitney U test (*wilcox.test*). P-values for mutation counts (indels and SNVs (**Fig. 2A**), 1-nt indels in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells (**Fig. 3A**), mutations in exonic vs non-coding regions in active and silent genes (**Fig. 5C-D**)) were calculated using two-tailed Fisher's exact test (*fisher.test*) and reported with odds ratio (O.R., ratio of ratios) and 95% condifence intervals (CI). O.R. values close to 1 indicate no difference in the ratios. Differences were determined statistically significant at a confidence level of 95%. Errors bars shown in **Fig. 3A** and **Fig. S4** are Sison and Glaz 95% multinomial confidence intervals from R package *DescTools* (0.99.25). Effect size reported for H3K36me3 signal in **Fig. 5E-F** was calculated using Cohen's d with Bessel's correction, implemented in R. Cohen's d values closer to 0 indicate smaller difference between two group means.

SUPPLEMENTAL MATERIALS AND METHODS

Analysis of thymic cell populations

Whole thymi of 4 *Mlh1*$^{-/-}$ and 4 *Mlh1*$^{+/+}$ mice were dissected and homogenized in HBSS + 2% FBS + 2 mM EDTA. Cells were frozen in HBSS + 10% DMSO + 40% FBS and stored at -80°C before analysis. Cells were thawed and transferred to warm HBSS + 2% FBS + 2 mM EDTA and washed twice. Live/dead staining was done for 15 minutes in +4°C with AmCyan (Invitrogen) 1:1000, followed by washing with HBSS + 2% FBS + 2 mM EDTA. Cells were stained for 45 minutes at +4°C for APC- TCRg/d (clone GL3, BioLegend) (1:400 dilution), AF780-TCRb (clone H57-597, BD Bioscience) (1:200 dilution), BV711-CD4 (clone RM4-5, BD Biosciences) (1:200 dilution), eVolve655-CD8a (clone 53- 6.7, eBioscience) (1:200 dilution), eFluor450-CD3 (clone 17A2, eBioscience) (1:150 dilution), PE- Cy7-NK1.1 (clone PK136, BD Bioscience) (1:200 dilution), BUV737-CD45.2 (clone 104, BD Horizon) (1:100 dilution) with Fc block CD16/32 (Invitrogen) (1:100 dilution) and rat serum (1:10 dilution). Cells were then washed with HBSS + 2% FBS + 2 mM EDTA, and resuspended to PBS. Cells were analyzed by flow cytometry with BD LSRFortessa and the results were further analyzed using FlowJo (10.5.3).

Huwe1 and *Mcm7* gene expression in T cell developmental sequence

Single-cell RNAseq gene expression data for hematopoietic stem cells (HSC), Slamf1 negative (-) and Slamf1 positive (+) multipotent progenitors (MPP), common lymphoid progenitors (CLP), pro T cells, immature T cells, and T cells for mm10 were downloaded from UCSC (http://genome.ucsc.edu/) as txt file tables containing sample IDs, category annotations and TMP values (Muris, 2018).

H3K36me3 fold change and deletion frequencies along RNApol2-positive and -negative gene bodies

Genes positive for RNApol2 were defined as described in Transparent Methods, under section "MMR-dependent mutation frequencies in 5′ and 3′ exons". Mean H3K36me3 fold change (FC) and deletion frequency in each single cell sample were calculated for each exon, and exon numbers along gene bodies were normalized to range from 0 – 1. Deletion frequencies were calculated by counting the absolute number deletions per each exon, and divided by the size of the exon. Deletion frequencies along gene bodies was plotted and smoothed using *glm*, and H3K36me3 FC using gam in *geom_smooth* from R package *ggplot2.*

Statistical testing

Percentages of different thymic cell types were tested using an implementation of two-sided permutation test that compares the difference of means to simulated distribution (Dermadi Bebek et al., 2014). Number of permutations used was 1000.

Supplemental Figures

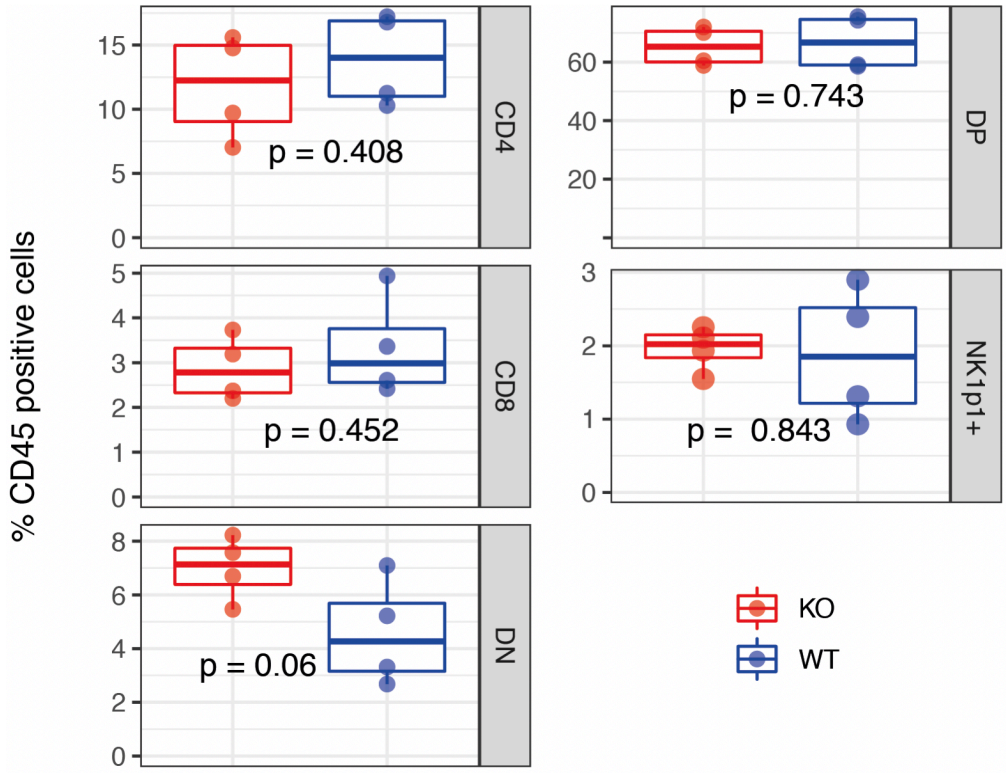


Figure S1 (related to Figure 1). *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ thymic lymphocytes.
CD4-CD8- double negative (DN), CD4+CD8+ double positive (DP), CD8+ single positive (CD8), CD4+ single positive (CD4), and TCRγδ T cells from *Mlh1*$^{+/+}$ (n = 4) and *Mlh1*$^{-/-}$ (n = 4) whole thymus. No significant difference was observed in different cell populations when comparing *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ mice. Data is shown as boxplots together with individual datapoints. P-values are two-sided values from permutation test (see Supplemental Methods section “Statistical testing”).

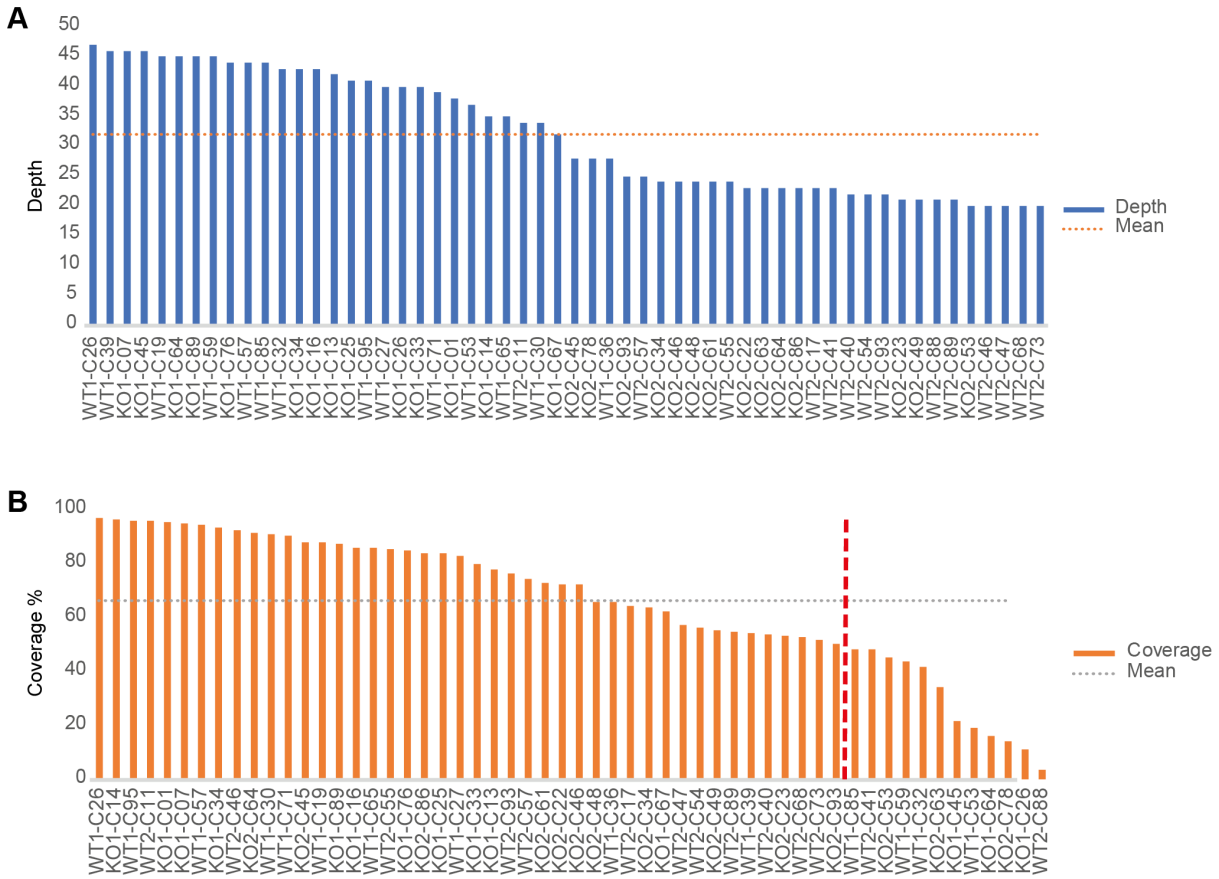


Figure S2 (related to Figure 1). Average depth and coverage of sequenced single T cell exomes.

(A) Average depth and (B) average coverage at depth ≥ 1 X in single T cells (n = 56). Dashed horizontal line marks mean depth or coverage across all sequenced T cells. Samples to the right of the vertical red dashed line in (B) were excluded from the mutational analysis.

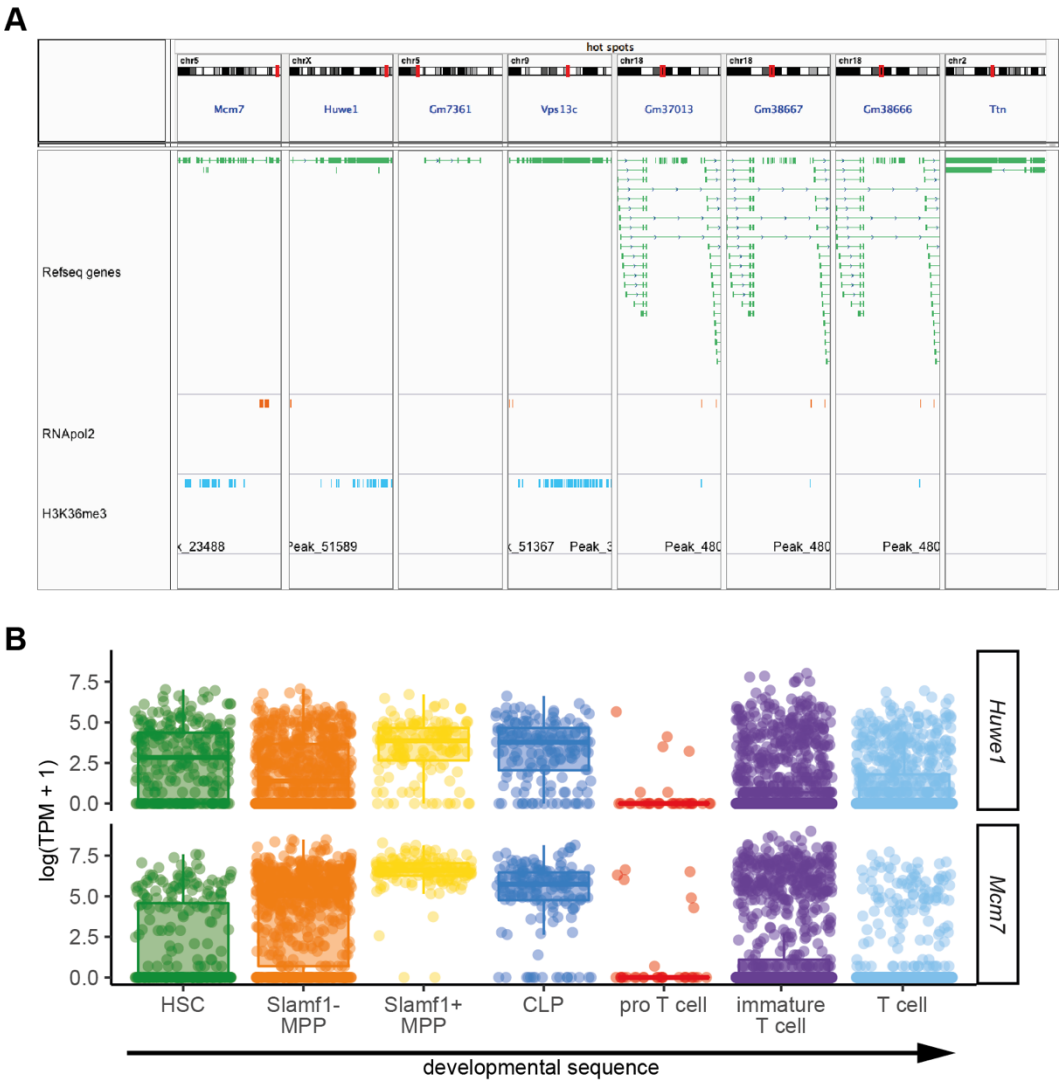


Figure S3 (related to Figure 2). H3K36me3 enrichment and transcriptional activity in mutational hotspot genes.

(A) RNA polymerase 2 and H3K36me3 ChIP-seq peak locations in mutational hotspot genes. *Mcm7*, *Huwe1* and *Vps13c* are positive (+) for, and *Gm7361* and *Ttn* are negative (-) for both RNA polymerase 2 and H3K36me3. *Gm37013*, *Gm68667* and *Gm38666* are inconclusive (-/+), because of the weak ChIP-seq signal. (B) Single-cell gene expression of *Mcm7* and *Huwe1* in developing T- lymphocytes. *Mcm7* and *Huwe1* are expressed from hematopoietic stem cells to naïve thymic T cells. Data is shown as boxplots together with individual datapoints.

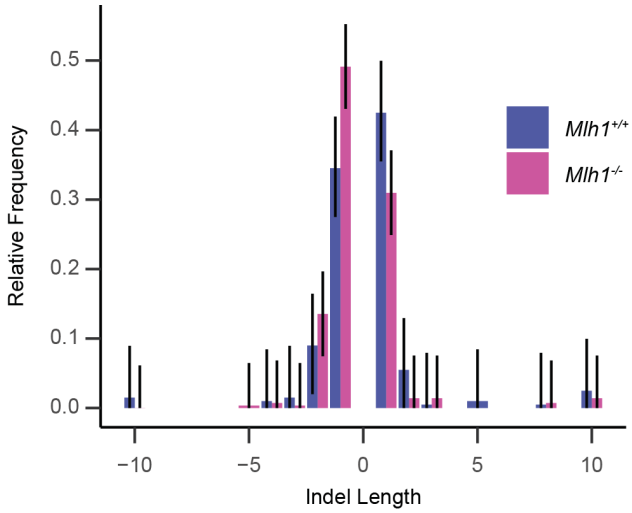


Figure S4 (related to Figure 3). Indel length distribution in bulk DNA samples.

Indel length distribution as relative frequencies with Sison and Glaz 95% multinomial confidence intervals of *Mlh1*$^{+/+}$ (n = 3) and *Mlh1*$^{-/-}$ (n = 3) bulk DNA samples, shown as relative frequencies. Bulk samples show the same trend in 1-nt indel bias as seen in single-cell samples. Technical replicates: 3, biological replicates: 2.

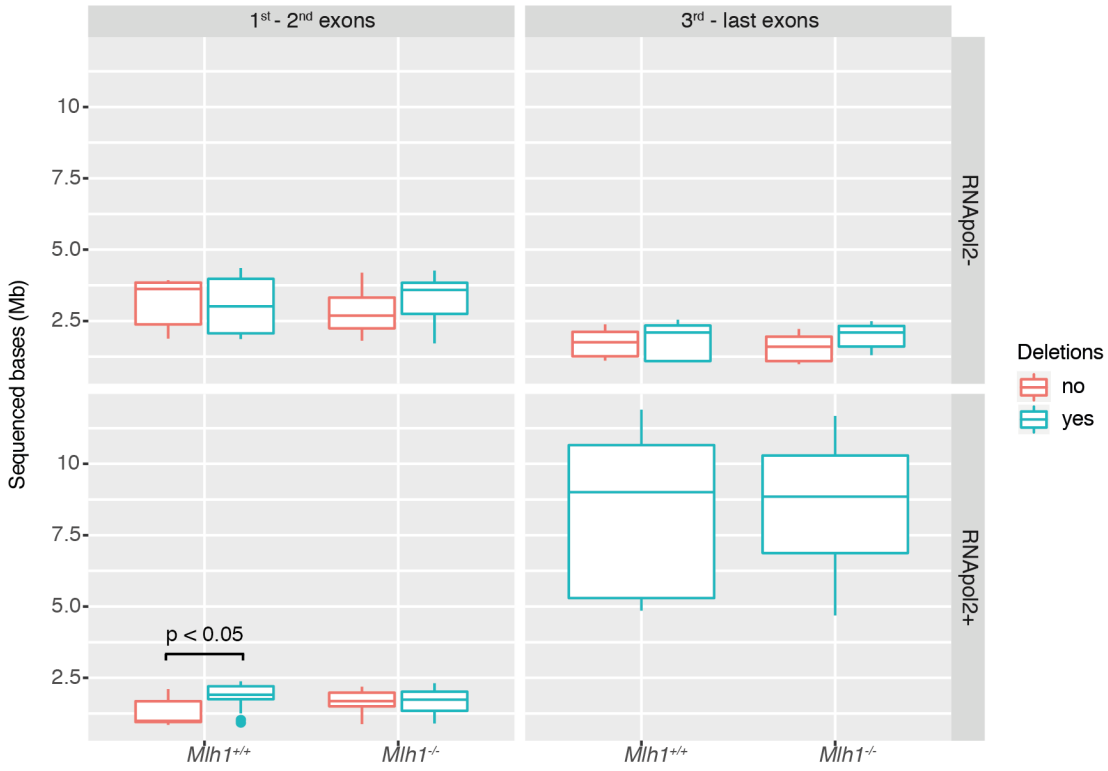


Figure S5 (related to Figure 5). Sequenced bases in 1st – 2nd exons and 3rd – last exons in RNApol2-negative (top panel) and -positive (bottom panel) genes.

Number of sequenced bases in 1st-2nd exons and 3rd-last exons shown as boxplots in 22 *Mlh1*$^{+/+}$ (three outlier cells included, see Methods) and 22 *Mlh1*$^{-/-}$ cells. A significant difference in sequenced bases between different exonic regions was seen only in RNApol2-positive 1st to 2nd exons of *Mlh1*$^{+/+}$ cells (p = 0.041, two-tailed Mann-Whitney U-test, n = 19 for *Mlh1*$^{+/+}$, n = 22 for *Mlh1*$^{-/-}$, outlier cells excluded – see Transparent Methods section "Outlier cells in single-cell exomes").

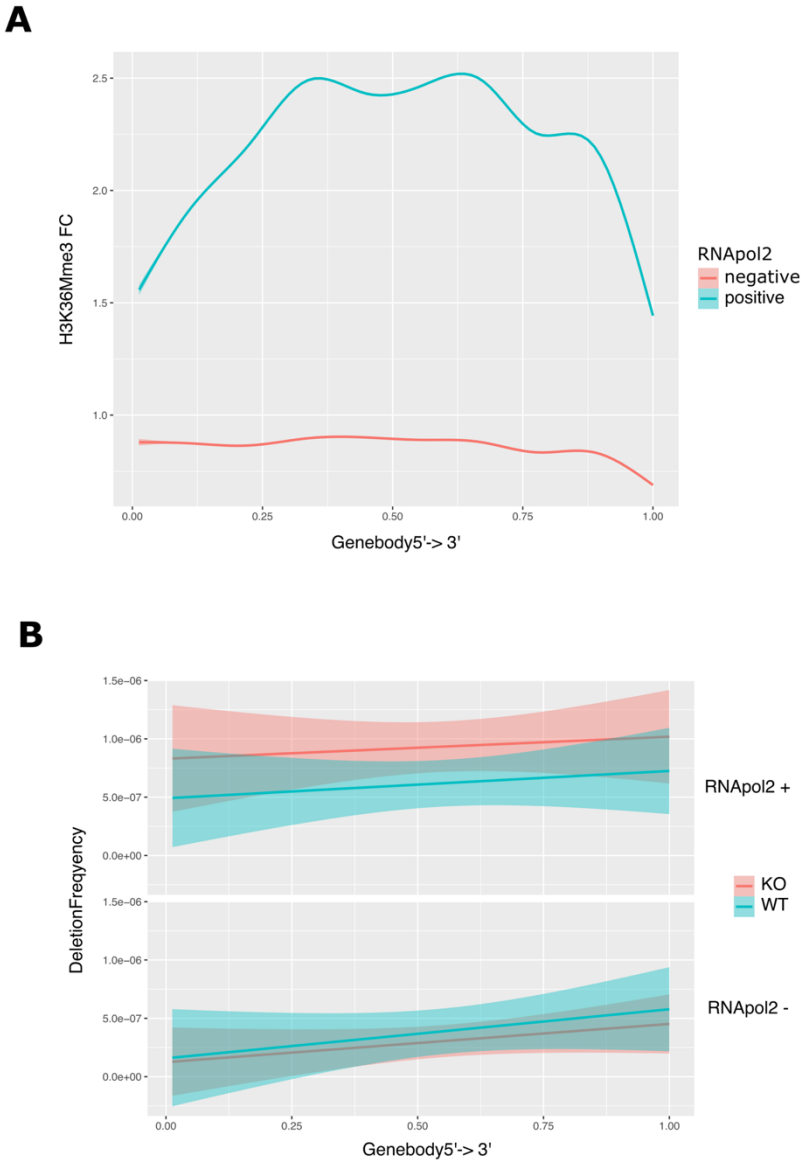


Figure S6 (related to Figure 5). H3K36me3 abundancy affects the deletion frequency along the gene body. A) H3K36me3 fold change along the gene body. RNApol2 positive genes have higher H3K36me3 fold change than RNApol2 negative genes. In the plot is shown generalized additive model (GAM) smoothing of H3K36me3 FC with 95% confidence interval B) Deletion frequencies in RNApol2 positive genes are higher in *Mlh1*$^{-/-}$ (KO) cells than in *Mlh1*$^{+/+}$ (WT), and increase from 5′ to 3′ end. In RNApol2 negative genes, no significant difference is observed between *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ cells. These results are complementary to the results shown in Figure 5 E-F. In the plot is shown GML smoothed deletion frequencies with 95% confidence intervals.

Table S1 (related to Figures 2-5). Mutation frequencies in *Mlh1*$^{+/+}$ and *Mlh1*$^{-/-}$ T cells.

Genotype	Region	Mutation type/ RNApol2 status	Mutation Frequency x10^{-7}			Figure
			Median	25% percentile	75% percentile	
Mlh1$^{-/-}$	3' exons	RNApol2-	3.73	0	5.92	5E-F
		RNApol2+	5.60	4.33	6.50	
	5' exons	RNApol2-	1.85	0	2.79	
		RNApol2+	3.83	2.29	7.40	
	coding	DEL	2.66	1.66	3.65	4A-C
		INS	2.70	2.08	4.55	
		SNV	19.2	9.64	21.4	
	exome	DEL	7.51	6.71	7.95	2C
		INDEL	13.7	12.7	15.2	2B
		INS	6.33	5.66	7.39	2C
		SNV	16.5	11.9	23.1	2B
	noncoding	DEL	7.23	5.23	9.05	4A-C
		INS	5.10	3.44	7.16	
		SNV	13.0	7.25	21.4	
	random	DEL	2.85	2.36	3.90	3C, E
		INS	1.36	0.467	2.20	
	repeat	DEL	44.2	34.0	50.8	
		INS	58.7	51.2	70.5	
Mlh1$^{+/+}$	3' exons	RNApol2-	1.97	0	4.36	5E-F
		RNApol2+	3.39	5.74	2.41	
	5' exons	RNApol2-	1.47	2.51	4.06	
		RNApol2+	2.27	0	2.83	
	coding	DEL	2.60	1.22	3.47	4A-C
		INS	2.73	2.02	4.16	
		SNV	11.5	7.43	3.07	
	exome	DEL	4.09	3.61	4.60	2C
		INDEL	9.22	7.84	11.9	2B
		INS	5.05	4.62	69.2	2C
		SNV	13.1	11.1	16.6	2B
	noncoding	DEL	4.17	3.28	4.48	4A-C
		INS	4.94	3.50	5.96	
		SNV	10.7	6.56	16.3	
	random	DEL	3.33	2.71	3.79	3C, E
		INS	0.638	0.235	1.38	
	repeat	DEL	11.2	8.93	13.2	
		INS	44.7	41.1	60.0	

Mutation frequencies in different genomic regions. Outlier cells (see Transparent Methods) were excluded from summary statistics.

SUPPLEMENTAL ACKNOWLEDGEMENTS

We would like to thank VA FACS Core Facility for access to BD LSRFortessa.

Supplemental References

AFGAN, E., BAKER, D., BATUT, B., VAN DEN BEEK, M., BOUVIER, D., CECH, M., CHILTON, J., CLEMENTS, D., CORAOR, N., GRUNING, B. A., GUERLER, A., HILLMAN-JACKSON, J., HILTEMANN, S., JALILI, V., RASCHE, H., SORANZO, N., GOECKS, J., TAYLOR, J., NEKRUTENKO, A. & BLANKENBERG, D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res, 46, W537-W544.

DERMADI BEBEK, D., VALO, S., PUSSILA, M., REYHANI, N., SARANTAUS, L., LALOWSKI, M., BAUMANN, M. & NYSTROM, M. 2014. Inherited cancer predisposition sensitizes colonic mucosa to address Western diet effects and putative cancer-predisposing changes on mouse proteome. J Nutr Biochem, 25, 1196-206.

EDELMANN, W., COHEN, P. E., KANE, M., LAU, K., MORROW, B., BENNETT, S., UMAR, A., KUNKEL, T., CATTORETTI, G., CHAGANTI, R., POLLARD, J. W., KOLODNER, R. D. & KUCHERLAPATI, R. 1996. Meiotic pachytene arrest in MLH1-deficient mice. Cell, 85, 1125-34.

FUNGTAMMASAN, A., ANANDA, G., HILE, S. E., SU, M. S., SUN, C., HARRIS, R., MEDVEDEV, P., ECKERT, K. & MAKOVA, K. D. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res, 25, 736-49.

KEANE, T. M., GOODSTADT, L., DANECEK, P., WHITE, M. A., WONG, K., YALCIN, B., HEGER, A., AGAM, A., SLATER, G., GOODSON, M., FURLOTTE, N. A., ESKIN, E., NELLAKER, C., WHITLEY, H., CLEAK, J., JANOWITZ, D., HERNANDEZ-PLIEGO, P., EDWARDS, A., BELGARD, T. G., OLIVER, P. L., MCINTYRE, R. E., BHOMRA, A., NICOD, J., GAN, X., YUAN, W., VAN DER WEYDEN, L., STEWARD, C. A., BALA, S., STALKER, J., MOTT, R., DURBIN, R., JACKSON, I. J., CZECHANSKI, A., GUERRA-ASSUNCAO, J. A., DONAHUE, L. R., REINHOLDT, L. G., PAYSEUR, B. A., PONTING, C. P., BIRNEY, E., FLINT, J. & ADAMS, D. J. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477, 289-94.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods, 9, 357-9.

LAWRENCE, M., GENTLEMAN, R. & CAREY, V. 2009. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics, 25, 1841-2.

LAWRENCE, M., HUBER, W., PAGES, H., ABOYOUN, P., CARLSON, M., GENTLEMAN, R., MORGAN, M. T. & CAREY, V. J. 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol, 9, e1003118.

LEUNG, M. L., WANG, Y., KIM, C., GAO, R., JIANG, J., SEI, E. & NAVIN, N. E. 2016. Highly multiplexed targeted DNA sequencing from single nuclei. Nat Protoc, 11, 214-235.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-9.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework fro analyzing next-generation DNA sequencing data. Genome Res, 20, 1297-1303.

TABULA MURIS CONSORTIUM. 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature, 562, 367-372.

OBENCHAIN, V., LAWRENCE, M., CAREY, V., GOGARTEN, S., SHANNON, P. & MORGAN, M. 2014. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics, 30, 2076-8.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841-2.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29, 308-311.